中国科学技术大学

本科毕业论文

Brain Tumor Segmentation on MRI with Missing Modalities using Multimodal Transformer

作者姓名：	BALKHANOV ALEKSANDR
学　　号：	PL20215001
专　　业：	计算机科学与技术
导师姓名：	王行甫
完成时间：	2024.05.24

ABSTRACT

Brain tumors exhibit high mortality and increasing incidence in recent years, creating significant challenges for timely diagnosis and effective treatment. Accurate segmentation of brain tumors on multimodal magnetic resonance imaging (MRI) is particularly crucial, as most normal tissues are unresectable during brain tumor surgery. This precision in segmentation is essential for delineating the boundaries of tumors, which can vary significantly in shape, size, and location, making treatment planning more effective and reducing the risk of harming critical brain structures.

In the past decade, the explosive development of artificial intelligence technologies has led to the presentation of a series of deep learning-based brain tumor segmentation methods that have achieved excellent results. These advancements have leveraged large datasets and sophisticated algorithms to enhance the accuracy and efficiency of tumor detection and segmentation.

In this thesis, I provide a comprehensive overview of existing methods for the segmentation of multimodal MRI brain tumor images, divided into four categories: traditional segmentation methods, segmentation methods based on classical machine learning, segmentation methods based on deep learning, and transformer-based methods. Each category is examined in detail to highlight the evolution of techniques and their respective strengths and limitations.

By identifying their advantages and disadvantages, we will delve into the most innovative approach to brain tumor segmentation, the multimodal transformer. This method represents a significant leap forward in the field, combining state-of-the-art technologies to overcome the challenges associated with incomplete data and varying imaging modalities.

Key words: Brain tumor segmentation, magnetic resonance imaging, multimodal transformer

摘　　要

脑肿瘤近年来表现出高死亡率和发病率的增加，给及时诊断和有效治疗带来了重大挑战。在多模态磁共振成像（MRI）上精确分割脑肿瘤尤为重要，因为在脑肿瘤手术中大多数正常组织是无法切除的。这种分割的精确性对于明确肿瘤的边界至关重要，因为肿瘤的形状、大小和位置可能有很大的差异，从而使治疗计划更加有效，减少损害关键脑结构的风险。

在过去的十年中，人工智能技术的爆炸性发展促使了一系列基于深度学习的脑肿瘤分割方法的出现，这些方法取得了卓越的成果。这些进展利用了大数据集和复杂的算法，增强了肿瘤检测和分割的准确性和效率。

在本论文中，我全面概述了现有的多模态MRI脑肿瘤图像分割方法，这些方法分为四类：传统分割方法、基于经典机器学习的分割方法、基于深度学习的分割方法和基于Transformer的方法。每一类方法都进行了详细审查，以突出技术的演变及其各自的优缺点。

通过识别它们的优势和劣势，我们将深入探讨脑肿瘤分割的最具创新性的方法——多模态Transformer。这种方法在该领域中代表了一个重大进步，结合了最先进的技术，克服了与不完整数据和不同成像模态相关的挑战。

关键词：脑肿瘤分割，磁共振成像，多模态变换器

Contents

ABSTRACT 2
摘要 3
First Chapter: Introduction 5
Section 1. Research Background and Significance 5
1.1 Multi-modal brain MRI 6
Section 2: Current State of Research 7
Second Chapter: Overview of methods 9
Section 1: Traditional segmentation methods 9
1.1 Threshold-based Segmentation 9
1.2 Region-based Segmentation 9
1.3 Segmentation Methods Based on Fuzzy Theory 10
1.4 Edge Detection-based Segmentation 11
Section 2: Traditional ML-Based Segmentation Methods 11
2.1 K-Nearest Neighbors 11
2.2 Support Vector Machine 12
2.3 Random Forest 12
2.4 Dictionary Learning 12
Section 3: Deep Learning-Based Segmentation methods 13
3.1 CNN-based segmentation models 13
3.2 FCN-based segmentation models 14
3.3 Encoder-Decoder segmentation methods 15
Section 4. Transformer-based methods 17
4.1. Pure transformer methods 17
4.2. Hybrid transformer methods 19
Third Chapter: Multimodal Transformer 23
Section 1: Method 23
1.1 Hybrid Modality-Specific Encoder 23
1.2 Modality-correlated Encoder 25
1.3 Convolutional Decoder 25
1.4 Auxiliary Regularizer 25
Fourth Part: Model Implementation 27
Section 1. Dataset and Implementation 27
Section 2. Experiments and Results 27
1.1 Comparison to Other Methods on Incomplete Multimodal Segmentation 27
Fifth Part: Conclusion and Outlook 30
Section 1. Conclusion 30
Section 2. Outlook 30
References 31
Acknowledgements 33

First Chapter: Introduction

Section 1. Research Background and Significance

The human brain, with its intricate network of neurons and synapses, stands as the pinnacle of complexity within the human body. Its functionality and structure are paramount to our understanding of life, cognition, and behavior. According to statistics from the World Health Organization (WHO), brain tumors represent a significant and growing health concern globally. These tumors encompass various types, including glioma, meningioma, pituitary adenoma, and nerve sheath tumors, with gliomas being the most prevalent and aggressive. These tumors, arising from glial cells due to a combination of genetic predispositions and environmental factors, pose severe health risks, often leading to debilitating symptoms such as seizures, headaches, and cognitive impairments.

Accurate localization and identification of brain tumors are paramount for effective diagnosis and treatment planning. The complexity of the brain's structure and the critical functions it performs necessitate a high degree of precision in medical interventions. Hence, the segmentation of brain tumors from medical imaging data plays a crucial role in enhancing diagnostic accuracy and prognostic capabilities in clinical practice. Over the years, extensive research efforts have been directed towards understanding the biological underpinnings of brain tumors and exploring innovative diagnostic and treatment approaches. These endeavors have led to significant advancements in tumor grading, genetic profiling, and therapeutic interventions. Such advancements have not only improved our understanding of tumor biology but also paved the way for personalized medicine approaches that can tailor treatment plans to the unique genetic and molecular profiles of individual tumors.

Additionally, imaging techniques such as Magnetic Resonance Imaging (MRI) have revolutionized the diagnosis and monitoring of brain tumors, enabling clinicians to visualize and analyze tumor characteristics with unprecedented clarity and precision. MRI-based imaging has become the cornerstone of brain tumor diagnosis, facilitating the development of minimally invasive surgical techniques and targeted treatment strategies. The ability to acquire high-resolution images of the brain’s internal structures allows for better planning of surgical interventions, ensuring that critical brain regions are preserved while maximizing tumor resection. MRI’s various sequences and modalities, including contrast-enhanced imaging, functional MRI, and diffusion tensor imaging, provide comprehensive information about tumor morphology, metabolism, and connectivity, which are crucial for effective treatment planning.

Central to these developments is the process of brain tumor segmentation, which involves delineating tumor regions from MRI images. This segmentation is critical for identifying tumor boundaries and distinguishing different tumor subtypes, thereby guiding treatment decisions and predicting patient outcomes. Accurate segmentation allows for precise mapping of tumor extent, facilitating targeted therapies such as radiation treatment and focused ultrasound. Moreover, it aids in monitoring treatment response and detecting recurrences, thereby playing a vital role in the longitudinal management of brain tumor patients. Advances in computational techniques, including machine learning and deep learning, have further enhanced the accuracy and efficiency of tumor segmentation, enabling automated and reproducible analysis of large imaging datasets. This integration of advanced imaging and computational methods holds the promise of significantly improving the prognosis and quality of life for patients with brain tumors.

1.1 Multi-modal brain MRI

As MRI technology evolves, the utilization of multi-modal MRI images for brain tumor segmentation becomes increasingly prevalent, offering a more comprehensive comprehension of both tumors and surrounding brain tissues. Specifically, MRI comprises four modalities: T1, T2, T1c, and FLAIR, each presenting supplementary imaging capabilities for diagnosing and monitoring brain tumors. Different MRI modalities furnish distinctive insights into tumor appearance and characteristics, thereby enhancing diagnostic precision (Table 1).

Modality	Property
T1	Emphasize the differences in T1 relaxation of the tissue. Brighter tissue indicates shorter relaxation time.
T2	Accentuate the tumor signal on T1 images.
T1c	Emphasize the differences in T2 relaxation (transverse) of the tissue. Brighter tissue indicates longer relaxation time.
FLAIR	Suppress the bright signal of cerebrospinal fluid. Highlight the small hyperintense tumor regions.

Table 1. The property of different modalities

For instance, the T1-weighted MRI modality furnishes high-resolution brain images, while the T2-weighted MRI modality offers insights into tissue fluid content, aiding in discriminating between tumor and normal brain tissue. Moreover, contrast-enhanced T1-weighted MRI reveals details about blood vessels and tumor enhancement patterns, facilitating tumor type classification and assessment of aggressiveness. Integrating multiple MRI modalities ensures a more exhaustive and accurate depiction of tumors and surrounding tissues, which is pivotal for effective glioma segmentation.

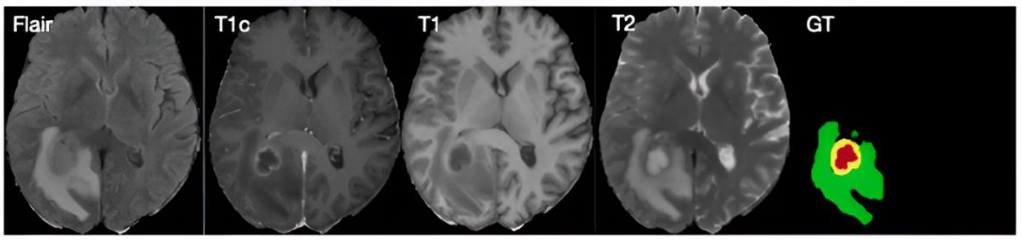

Fig 1. Depiction of four MRI modalities

We can see in Fig. 1 an example of the aforementioned four MRI modalities and the true regions of brain tumors from left to right: Flair, T1c, T1, and T2. These images are shown alongside the ground truths (GT) of brain tumors, with necrotic and non-enhancing tumors depicted in green, edema in red, and enhancing tumors in yellow. This comprehensive representation allows us to observe the distinctive contributions of each MRI modality to tumor visualization. Specifically, it can be observed that T2 and Flair provide useful information about the entire tumor, capturing the overall tumor mass and its surrounding effects. In contrast, T1 and T1c are more effective in highlighting the tumor core, which includes the necrotic, non-enhancing, and enhancing tumor regions. This detailed view of different tumor components is crucial for accurate diagnosis and effective treatment planning, as it helps in distinguishing between the various pathological features within the brain. The integration of these modalities enhances our ability to assess the extent and nature of brain tumors comprehensively.

Section 2: Current State of Research

Over the past decade, the rapid development of artificial intelligence and deep learning technologies has significantly advanced the field of medical image analysis. Deep learning models, particularly convolutional neural networks (CNNs) and more recently transformer-based architectures, have shown superior performance in image segmentation tasks. These models leverage large datasets and sophisticated algorithms to enhance the accuracy and efficiency of brain tumor segmentation, offering substantial potential for improving clinical workflows (Fig. 2)

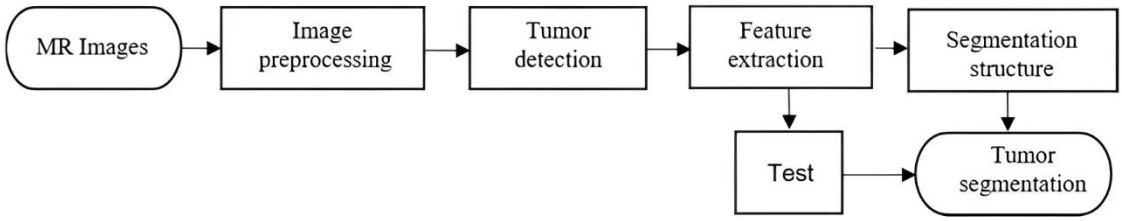

Fig 2. Tumor segmentation process flowchart

Brain tumor segmentation includes four main categories of methods: traditional segmentation methods, segmentation methods based on classical machine learning, segmentation methods based on deep learning, and transformer-based methods (Fig. 3)

Traditional segmentation methods include algorithms primarily based on image intensity analysis and geometric characteristics. These methods often involve thresholding, which determines an intensity threshold value to separate pixels into classes (tumor and healthy tissue), region growing methods, where a region is gradually expanded by including neighboring pixels that meet certain criteria, active contours, which adjust a contour under the influence of internal and external forces to find the boundary between different image regions and clustering methods such as k-means, which group pixels into clusters based on their intensity. The advantages of these methods are their simplicity of implementation and interpretation, but they are sensitive to noise and dependent on initial conditions.

Segmentation methods based on classical machine learning involve algorithms that learn from pre-labeled data to classify pixels. These are feature-based methods, including the extraction of various features from the image, such as textures, shape, and intensity, and using them to train classifiers like SVM (Support Vector Machine) or random forests, hidden Markov models that consider spatial dependencies in the image and Bayesian classifiers, which determine the probability that a pixel belongs to a particular class based on its features. The advantages of these methods lie in the ability to consider various image features, but they require careful feature selection and substantial labeled datasets for training.

Segmentation methods based on deep learning use neural networks to automatically extract features and classify pixels. The primary tool for image segmentation is convolutional neural networks (CNN), capable of automatically extracting complex spatial and textural features. The U-Net architecture, specifically designed for biomedical image segmentation, includes encoding and decoding paths with skip connections. Also used are multi-level and multi-scale networks, which enhance segmentation by using information from different levels and scales of the image. The advantages of these methods include

high accuracy and the ability to automatically extract features, but they require significant computational resources and large amounts of data for training.

Transformer-based segmentation methods utilize transformer architectures, originally developed for natural language processing tasks. There are two main types of transformer-based methods: pure transformer methods and hybrid transformer methods. Pure transformer methods rely entirely on transformers to process and segment the images. Hybrid transformers combine the capabilities of CNNs for local feature extraction and transformers for modeling global dependencies. Multimodal transformers can process and integrate information from different MRI modalities (T1, T2, FLAIR, etc.), providing a more comprehensive representation of tumors. The advantages of these methods include the ability to model long-term dependencies and integrate data from multiple sources.

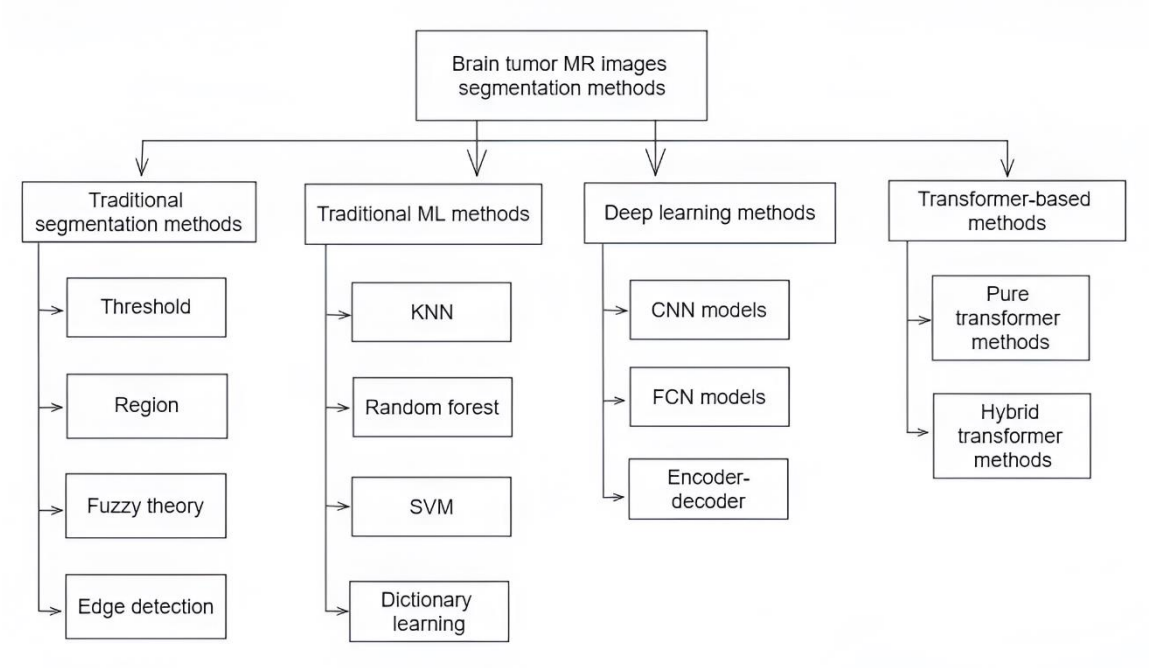

Fig 3. Main methods of Brain tumor segmentation

In this thesis, we will conduct an in-depth examination and analysis of the principal methods for brain tumor segmentation, highlighting their strengths and weaknesses. This examination will cover a range of methodologies from traditional segmentation techniques to more contemporary approaches, providing a comprehensive understanding of their evolution and application in medical imaging. Following this, we will transition to discussing the most advanced method to date: the multimodal transformer. We will delve into its functionality, examining how it integrates different modalities to provide a more accurate and holistic view of brain tumors. Additionally, we will implement this method on a dataset BraTS 2020, rigorously testing its performance and adaptability. This practical application will enable us to draw informed conclusions about the future prospects of brain tumor segmentation, considering the potential for further advancements and integration into clinical practice. Through this comprehensive study, we aim to contribute valuable insights and pave the way for more effective and reliable diagnostic tools in the field of neuroimaging.

Second Chapter: Overview of methods

Section 1: Traditional segmentation methods

Traditional segmentation methods can generally be categorized into four groups based on different theories and focuses: threshold-based segmentation, region-based segmentation, segmentation based on fuzzy theory, and edge detection-based segmentation (Fig. 4)

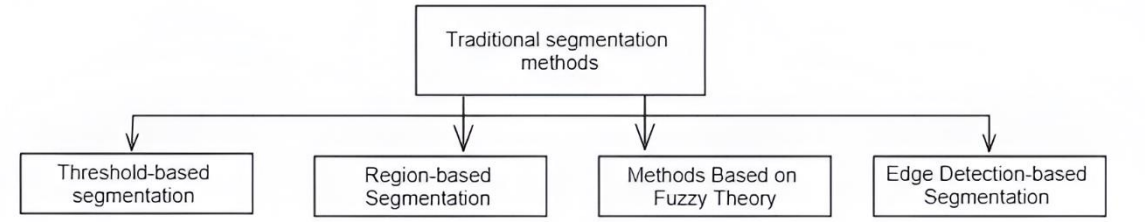

Fig 4. Groups of traditional segmentation methods

1.1 Threshold-based Segmentation

Threshold-based segmentation stands out as one of the simplest approaches. It starts with the assumption that pixels falling within a specific range belong to identical categories. By appropriately setting a threshold, brain tumor images can be partitioned into target and background regions. Moreover, employing various thresholds allows for the division of tumors into multiple regions. Over time, through continuous research efforts, an enhanced threshold segmentation algorithm emerged, substantially elevating the accuracy of this segmentation technique. This advancement is attributed to the method's ability to heighten noise sensitivity by leveraging local pixel neighborhood information. The most adaptive threshold segmentation method effectively addresses issues related to uneven gray levels, enhances image contrast, and significantly improves segmentation metrics such as DSC and JS for MRI images of brain tumors [1]. While threshold-based segmentation is relatively straightforward, the effectiveness of its outcomes heavily relies on the chosen threshold's magnitude, emphasizing the critical importance of threshold selection. Furthermore, this method's applicability is restricted to simpler images and may encounter challenges when applied to more complex ones.

1.2 Region-based Segmentation

Among the most common region-based segmentation methods, the watershed algorithm and the region-growing algorithm stand out. The watershed algorithm is based on the principles of mathematical morphology. It compares the processed image to a geographic terrain, where the gray pixel values represent terrain elevation. Local minima and their surrounding areas are identified as catchment basins, assuming the presence of water-permeable holes in each local minimum. As these basins gradually fill with water, they eventually reach peak levels, at which point the process concludes. Meanwhile, dams, called watersheds, block the water flow between stagnant basins (Fig. 5). For instance, there is a watershed segmentation method that uses internal or external markers to compute the morphological gradient of both the input and original images, followed by applying a watershed transform to obtain segmentation results. Another method uses a hybrid approach that combines threshold segmentation with watershed segmentation, yielding

segmentation results that surpass those achieved by either method alone [2]. While the watershed-based segmentation algorithm offers advantages such as obtaining closed contours and providing contour information for subsequent processing, it is susceptible to noise interference, which can lead to over-segmentation.

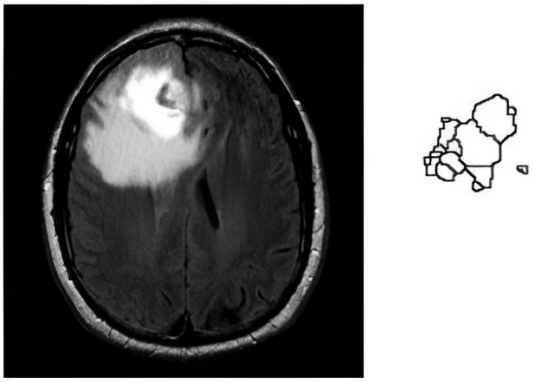

Fig 5. The example of watershed algorithm segmentation on brain MRI

On the other hand, the region-growing algorithm involves aggregating all pixel points that meet a specific criterion into a single region, thereby achieving pixel segmentation. Notably, this segmentation method requires each pixel to be included in a particular region, with interconnected pixels in the region adhering to predefined similarity conditions. Additionally, it ensures that different regions remain disjoint and possess distinct properties. For example, an automatic seed region growth method automates the assignment of seed values, eliminating the drawbacks of manual interaction and enhancing image segmentation efficiency. Region-based segmentation methods are characterized by their computational simplicity and high accuracy, especially in extracting high-quality regional features, making them suitable for segmenting small objects. However, they are sensitive to noise and can result in perforations within the segmented region.

1.3 Segmentation Methods Based on Fuzzy Theory

Fuzzy theory-based segmentation methods have also gained significant recognition. Researchers have employed the fuzzy C-means (FCM) clustering algorithm, widely used for segmenting MRI brain tumor images. The developed K-FCM method resulted from combining the FCM algorithm with the K-means algorithm. Comparative experiments demonstrated that K-FCM, compared to FCM, has higher precision in segmenting MRI brain tumor images while reducing computational complexity. In another study, a fuzzy C-means algorithm based on fuzzy C-means and a neutral set was proposed. This algorithm introduces a distance constraint in the objective function to address the issue of insufficient prior knowledge and achieve satisfactory segmentation results [3]. Furthermore, a level-set method based on fuzzy superpixel clustering and the Boltzmann grid method was proposed, which automatically segments brain tumors and exhibits robustness to image intensity and noise.

Fuzzy theory-based segmentation methods effectively address issues such as incomplete image information and fuzziness. They exhibit strong compatibility and can be synergistically employed with other methods. However, due to significant computational requirements and high temporal complexity, they face challenges when dealing with large-scale datasets.

1.4 Edge Detection-based Segmentation

The principle of edge detection-based segmentation and contour delineation involves segmenting the target area based on the acquired edges and then outlining the contour of the target area. For example, the FCM algorithm was integrated into an active contour model. The automatic selection of the initial contour of the model using FCM reduces the level of human-computer interaction [4]. Additionally, enhancements were made to address the issue of blurred edge contours and uneven intensity in MRI images. Compared to other traditional segmentation methods, edge detection-based segmentation methods focus on edge information in the image and integrate edges into contours, demonstrating higher noise resistance. However, noise resistance is inversely proportional to accuracy; therefore, better noise suppression leads to reduced accuracy. Conversely, improving accuracy will decrease noise resistance.

Section 2: Traditional ML-Based Segmentation Methods

Methods for segmenting MRI brain tumor images using traditional machine learning rely on predefined features to train classification models. Typically, they are categorized into two tiers: organizational and pixel levels. At the organizational tier, the classifier must ascertain the organizational structure type each feature pertains to, while at the pixel level, it determines the category each pixel falls into. Essential algorithms in traditional machine learning encompass k-nearest neighbors (KNN), support vector machine (SVM), random forest (RF), dictionary learning (DL), and similar methods (Fig. 6)

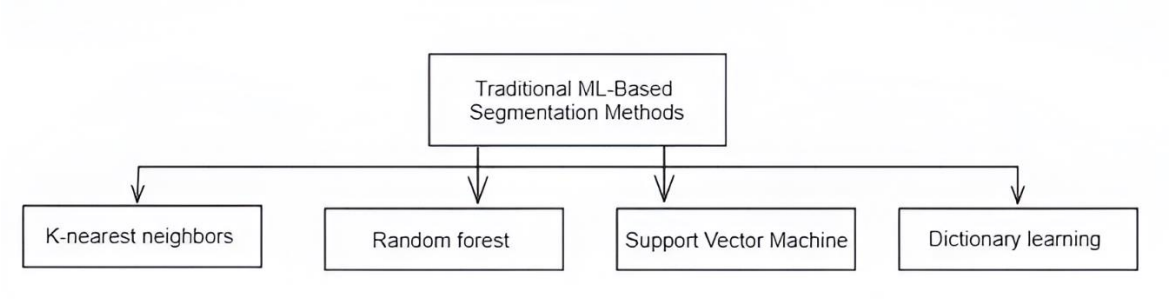

Fig 6. Traditional ML-based segmentation methods

2.1 K-Nearest Neighbors

K-Nearest Neighbors is a simple and intuitive method used for brain tumor segmentation. It classifies each pixel or voxel by comparing it with its k-nearest neighbors in the feature space. In one model applying this method, researchers treated each brain image as an individual dataset and employed the KNN algorithm for segmentation, achieving high accuracy within a minute per brain image, thus enhancing segmentation efficiency [5]. In another method, images were initially segmented using the Otsu method and K-means clustering, then features were extracted using discrete wavelet transformation, and feature dimensionality was reduced using PCA to enhance features for SVM classification. Generally, the KNN method is easy to implement and provides relatively high accuracy, but it can be computationally intensive, especially with large datasets.

2.2 Support Vector Machine

Support Vector Machine is a powerful supervised learning algorithm that finds the optimal hyperplane separating different classes in the feature space. SVM is used for brain tumor segmentation by considering changes in signal intensity characteristics and other features of MRI brain tumor images. Some studies applied a similar approach using the Proximal Support Vector Machine (PSVM), which simplifies the original convex quadratic programming problem by solving primary linear equations. The PSVM method proved more accurate than SVM in segmenting MRI brain tumor images. Additionally, there is a method that first segments the image into super-voxels, then segments the tumor using MRF while simultaneously estimating likelihood functions and extracting features using a multilevel wavelet filter [6]. SVM offers strong generalization capabilities and relatively simple computation but requires careful parameter selection and choice of kernel functions.

2.3 Random Forest

RF is an ensemble learning method that builds multiple decision trees during training and merges their results to improve accuracy and control overfitting. It is effective in segmenting brain tumors by leveraging multiple features, such as intensity, gradient, and entropy. For example, an automatic segmentation algorithm based on RF includes several significant characteristics, such as image intensity, gradient, and entropy, to create multiple classifiers for pixel classification in multispectral MRI brain images, combining results for segmentation outcomes [7]. In some studies, researchers adjusted the weight of misclassified samples and reduced the weight of correctly classified samples during training. The classifier then assigned new weights to the samples to ensure a positive correlation between the weights of all decision trees and their classification ability. Ultimately, the enhanced RF consisted of two parts: image intensity characteristics and original image characteristics, extracted via curve and wavelet transformation. Researchers examined the correlation between image minimization characteristics and image attributes, effectively selecting characteristics, and then classifying characteristics in multimodal MRI images using RF. Compared to the RF algorithm alone, the proposed method simultaneously improved DSC, PPV, and TPR measures. Then, researchers trained a specific random forest classifier using semi-supervised learning, treating image segmentation as a classification task, and effectively combining preoperative and postoperative MRI images to improve brain tumor segmentation after surgery [8]. In summary, RF is robust to noise and capable of handling high-dimensional data, but its performance can be limited by the quality of the individual trees.

2.4 Dictionary Learning

Dictionary learning is a learning method that simulates dictionary search. The dictionary itself is set as a dictionary matrix, and the method utilized is a sparse matrix. The dictionary search process is achieved by multiplying the sparse matrix by the dictionary matrix, and then optimizing the dictionary matrix and sparse matrix to minimize the error between the found value and the original data. Researchers converted the superpixel feature into high-dimensional space. Depending on different error values of various areas when modeling the brain tumor dictionary, brain tumor MRI image segmentation was conducted, enhancing segmentation accuracy. A notable approach is the multiscale fuzzy learning algorithm based on dictionary learning. This algorithm effectively combines dictionary learning with fuzzy algorithms and fully considers differences between the target area and the background, as well as consistency within the

target area. This method can describe the gray and texture characteristics of different image areas and rapidly and accurately segment the image. DL effectively reduces computational complexity and accelerates processing speed, but it requires a well-constructed dictionary and can be sensitive to the quality of input data.

Traditional machine learning algorithms demonstrate superior algorithmic performance compared to many traditional segmentation methods. However, they exhibit several limitations when applied to segmenting MRI brain tumor images. For instance, the KNN algorithm offers ease of implementation and relatively high accuracy in predicting brain tumor regions, albeit with considerable computational requirements. Similarly, the support vector machine relies on a robust theoretical framework and is determined by multiple support vectors, boasting relatively simple computation and strong generalization ability but demanding meticulous parameter selection and kernel function choice. Random forest addresses over-segmentation effectively, accommodating diverse data types and showcasing robust noise resistance. While capable of parallel processing to reduce processing time, it may adversely affect the analysis of low-dimensional tumor data. The dictionary learning approach, akin to dimensionality reduction, streamlines computational complexity and accelerates processing speed, yet imposes heightened prerequisites for tumor data.

Section 3: Deep Learning-Based Segmentation methods

With different network structures in mind, brain MR image segmentation methods are rooted in deep learning and can be categorized into those leveraging Convolutional Neural Networks (CNNs) for general brain MR image segmentation, and those utilizing Fully Convolutional Networks (FCNs) for brain tumor MR image segmentation, as well as the encoder-decoder approach for brain MR image segmentation (Fig.7)

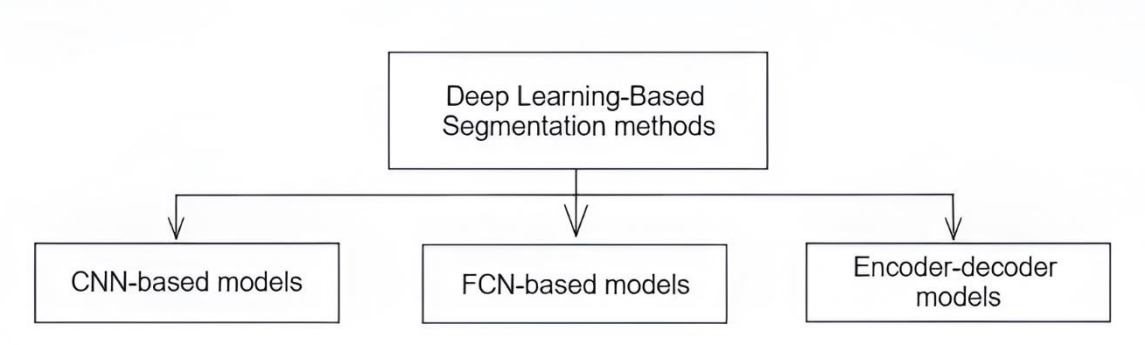

Fig 7. DL-based segmentation methods

3.1 CNN-based segmentation models

Convolutional neural networks (CNN) have become one of the primary methods for image analysis, including the segmentation of brain MR images. The main idea of CNNs is to use convolutional layers for automatic feature extraction from input data (Fig. 8). These layers include convolutions, pooling, and activation functions such as ReLU. Convolutional layers apply filters that scan the image and detect various patterns, such as edges and textures. Pooling layers reduce the dimensionality of the data while preserving important features.

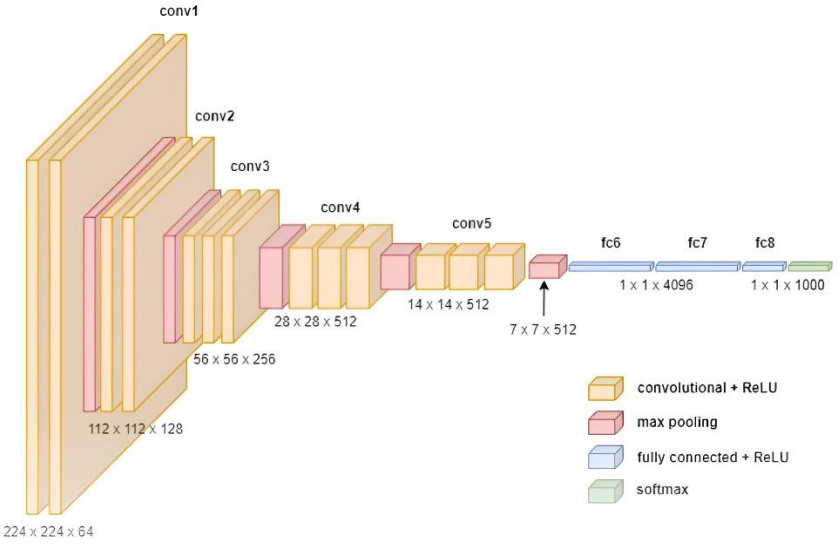

Fig 8. Convolutional neural network (CNN) architecture

These networks directly process images, autonomously extracting features and demonstrating high resistance to image translation, scaling, and other variations. In recent years, various network models based on CNNs, including Network in Network, VGG, GoogleNet, Res-Net, and others, have been widely used for medical image segmentation [8]. Among them, VGG networks excel in feature extraction and ensure convergence with a minimal number of training iterations. However, increasing network depth can lead to issues such as gradient problems beyond a certain depth threshold. To overcome network degradation, deep residual networks (ResNet) have been proposed, showing promising results in segmentation tasks. Researchers have explored techniques such as combining 3D convolutional neural networks with dense connections, pre-training the model, and initializing it with acquired weights to enhance performance in brain tumor MR image segmentation tasks.

Also noteworthy is a more hierarchical convolution-based neural network (Multi-Cascaded Convolutional Neural Network, MCCNN) combined with fully connected conditional random fields (CRFs) for brain tumor segmentation. Initially, rough segmentation by a multi-classification convolutional neural network is followed by fine segmentation through fully connected random fields to achieve batch segmentation and improve accuracy [9]. Despite the ability of CNN-based segmentation algorithms to autonomously extract features and handle high-dimensional data, they may experience information loss during pooling, leading to reduced interpretability.

3.2 FCN-based segmentation models

Fully Convolutional Networks (FCNs) are an extension of CNNs optimized for segmentation tasks. Unlike standard CNNs, FCNs replace fully connected layers with convolutional layers, allowing the network to process input images of arbitrary sizes and produce segmentation maps of the same dimensions. This makes FCNs ideal for tasks requiring dense prediction maps, such as brain tumor segmentation (Fig. 9).

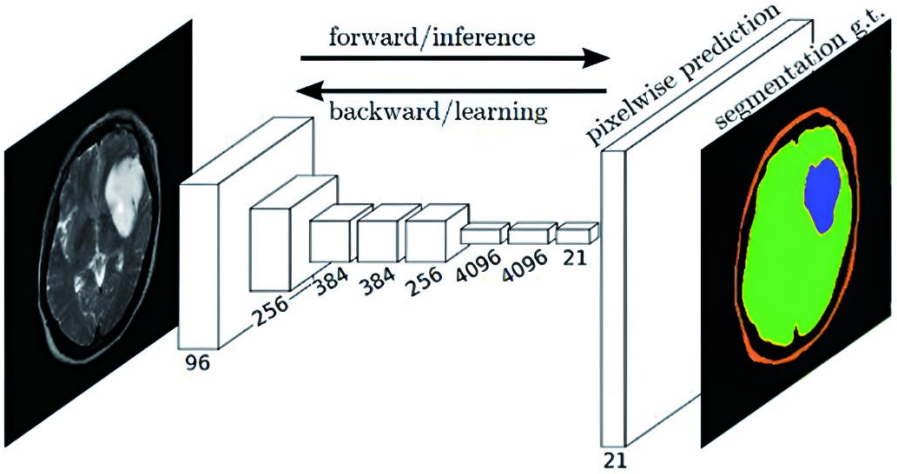

Fig 9. FCN-based network structure for brain tumor segmentation

Compared to pixel-level classification, image-level classification and regression tasks, both of which aim to derive a probable value for image classification, are more suitable for utilizing the CNN structure. For the semantic segmentation of images, FCNs prove to be more effective. FCNs do not impose any restrictions on the size of the input image. Instead, an upsampling process occurs at the last convolution layer. This process ensures the prediction for each pixel while preserving spatial information, thus achieving pixel-level classification. Essentially, FCNs function as a method to classify and segment images at the pixel level. Therefore, the semantic segmentation model based on FCNs aligns more closely with the requirements of medical image segmentation.

Although the segmentation algorithm based on FCNs can predict the category of each pixel and transform the image classification level to the semantic level while retaining positional information in the original image, it has drawbacks such as low computational efficiency, high memory consumption, and a relatively small receptive field.

3.3 Encoder-Decoder segmentation methods

The use of encoder-decoder structures has become a prevalent approach in segmenting brain tumor MR images, offering a reliable framework for accurately delineating tumor boundaries. At the core of this architectural paradigm lie two key components: the encoder and the decoder. The encoder component processes input images through a neural network, extracting corresponding feature maps that contain important information indicative of the presence of a tumor. In contrast, the primary function of the decoder is to semantically interpret these feature maps, assigning categorical labels to individual pixels to effectively delineate tumor boundaries. In the realm of encoder-decoder-based segmentation methods, the architecture of encoder networks typically draws inspiration from proven classification models such as VGG. By adapting these architectures for segmentation tasks, researchers aim to leverage existing weight parameters obtained from extensive training on large datasets. Consequently, the design of the decoder becomes a critically important factor in the overall network performance and segmentation efficiency.

One notable example in this field is the SegNet model (Fig. 10), distinguished by its considerable architectural depth compared to its contemporaries [11]. Utilizing a 13-layer VGG-16 network in its encoder module, SegNet excels in preserving important spatial information during the encoding phase. Subsequently, in the decoding phase, the model employs sophisticated up-sampling techniques to refine low-resolution input features, thereby ensuring precise segmentation results.

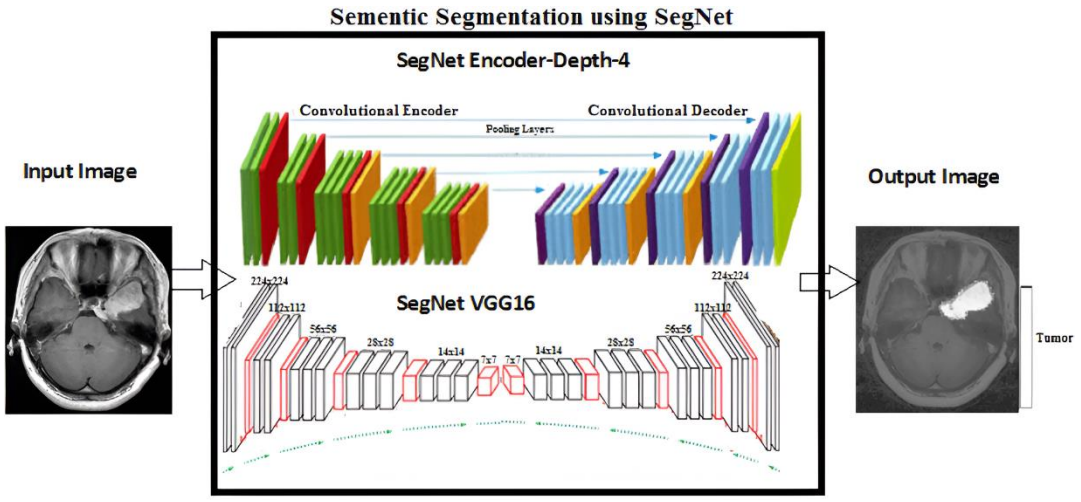

Fig 10. Pipe-line of Brain Tumor segmentation based on SegNet

Another significant contender is the U-Net model, a major contribution to the field of Fully Convolutional Networks (FCN), which has received widespread acclaim for its efficacy in brain tumor segmentation tasks (Fig. 11). Characterized by its unique encoder-decoder architecture, U-Net employs skip connections to facilitate the seamless transfer of features from the encoding path to their corresponding positions in the decoding path. This mechanism enables the model to capture detailed contextual information, which is critically important for accurate segmentation [12].

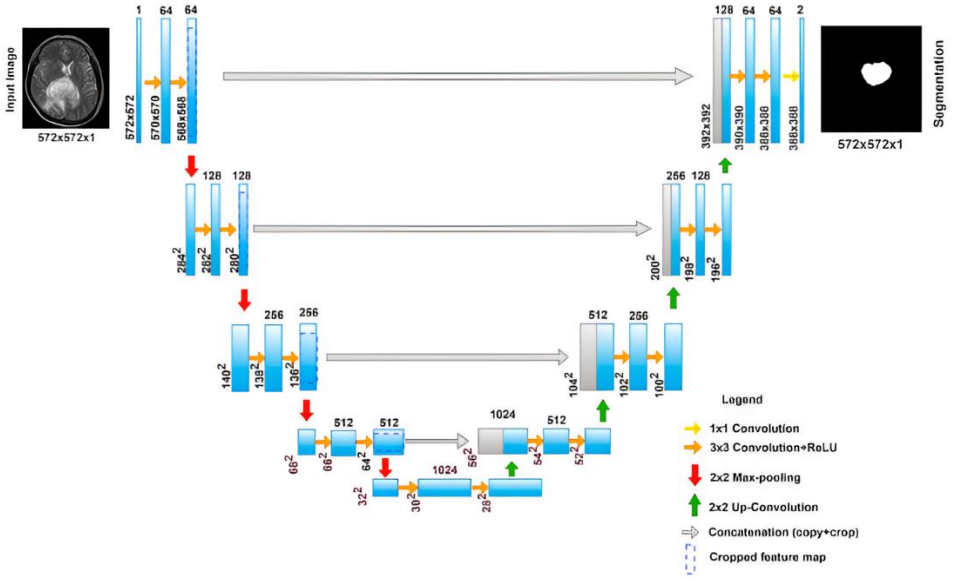

Fig 11. U-Net structure for brain tumor segmentation.

Despite the promising capabilities of encoder-decoder-based segmentation algorithms for seamlessly merging high-resolution and low-resolution information and capturing features across multiple scales, an existing limitation lies in the insufficient connectivity between the encoding and decoding processes.

Section 4. Transformer-based methods

Transformer-based methods for multi-modal brain tumor segmentation have garnered attention due to their potential to provide accurate and comprehensive segmentation results, crucial for effective diagnosis and treatment planning. While traditional neural networks (CNNs) have traditionally been the dominant approach for brain tumor segmentation, they are inherently suboptimal in capturing non-local feature relations, often requiring the construction of complex and deep network architectures.

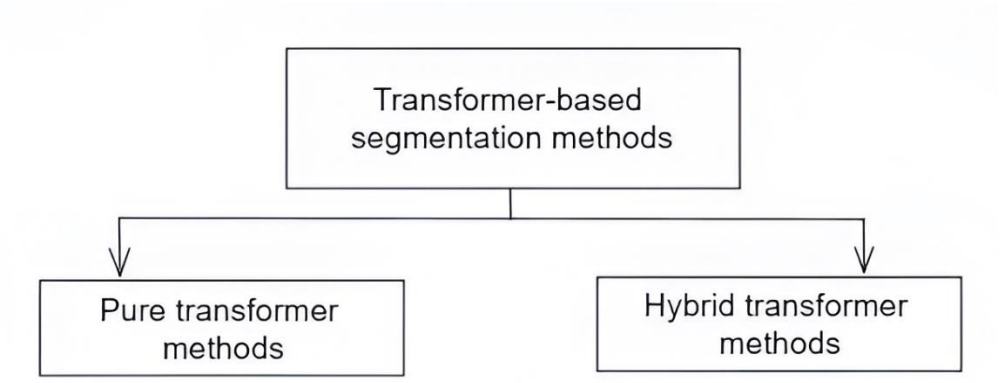

Fig 12. Two main categories of Transformer-based methods

In contrast, segmentation methods based on transformers offer a novel approach by considering all local image patches to establish global feature relations. This comprehensive approach enables better exploration and modeling of correlations between tumor regions and backgrounds, thereby enhancing brain tumor segmentation accuracy. Presently, a variety of transformer-based methods have emerged for brain tumor segmentation, showcasing promising performance in this domain. These methods typically adhere to a core principle, leveraging transformers with global receptive fields and innovative designs to enhance their ability to model local feature relations. By doing so, transformers can effectively capture both global and local information, contributing to improved segmentation outcomes. Based on transformer architectures, these methods can be classified into two main categories: pure transformer methods and hybrid transformer methods (Fig.12).

4.1. Pure transformer methods

In this section, we delve into pure transformer methods, which rely solely on transformer-based blocks without integrating CNNs. These methods represent a significant departure from traditional approaches, leveraging the power of transformers like ViT and SwinT, prevalent in computer vision [13].

The SwinT transformer has become a popular choice as the foundational transformer architecture for brain tumor segmentation in numerous studies. For example, BTSwin-UNet is a notable model that employs a symmetrical network based on 3D SwinT in a U-shaped encoder-decoder structure (Fig.13). This network uses 3D SwinT blocks as the primary building units for creating the encoder-decoder framework with skip connections. Feature manipulation, such as up- and down-sampling, is achieved through convolutions and deconvolutions, respectively. Moreover, BTSwin-UNet incorporates a self-supervised learning mechanism. This approach involves randomly masking certain regions within input images and subsequently using an additional decoder to restore these masked regions.

Using the masked autoencoder as a powerful feature extractor, this self-supervised learning scheme effectively encourages the encoder to extract high-quality features. As a result, BTSwin-UNet demonstrates superior performance in brain tumor segmentation compared to alternative methods.

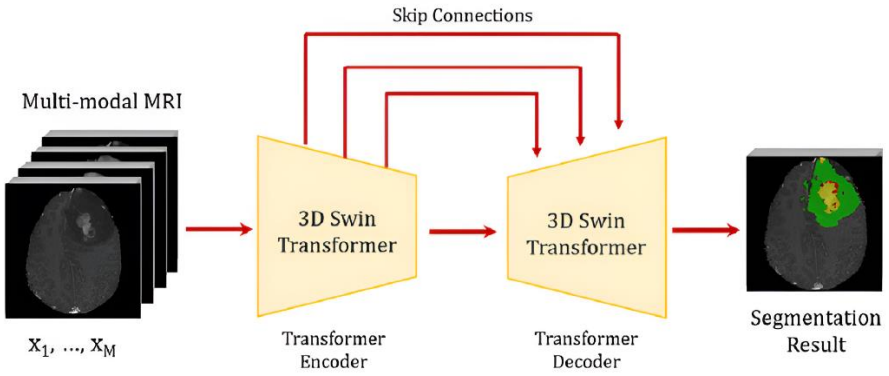

Fig 13. BTSwin-UNet model

Subsequently, researchers proposed CR-Swin2-VT (Fig.14), using parallel CSwinT and SwinT blocks in the encoder to enhance feature extraction quality and improve brain tumor segmentation outcomes. Additionally, CR-Swin2-VT introduces a feature fusion mechanism using learnable weights to merge the outputs of CSwinT and SwinT blocks at each stage [14]. These approaches effectively harness the advantages of SwinT, capturing local feature relations via memory-efficient window self-attentions while establishing global feature relations through shifted windows, resulting in satisfactory brain tumor segmentation outcomes.

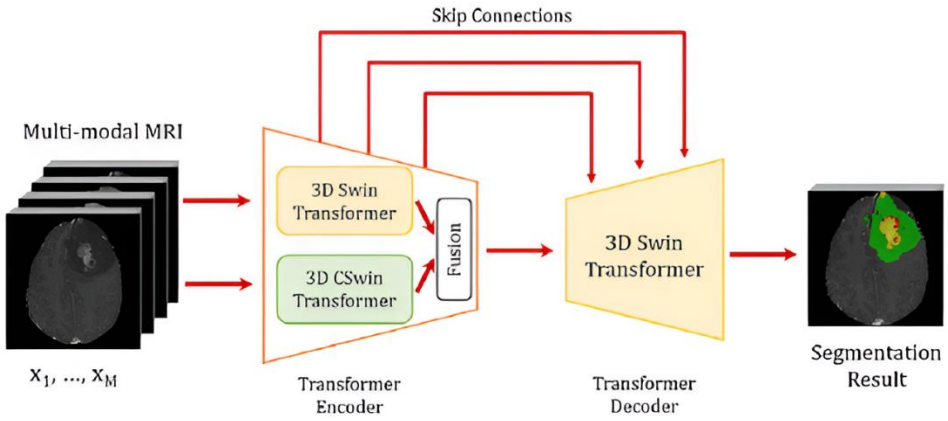

Fig 14. CR-Swin2-VT model

Following the HRNet paradigm, a High-Resolution Swin Transformer Network (HRSTNet) was developed using standard SwinT blocks (Fig.15). Unlike the conventional U-shaped encoder-decoder architecture, HRSTNet employs a parallel multi-resolution path to retain high-resolution information within each network stage. Although this approach increases computational demands, it simultaneously mitigates the loss of spatial information caused by feature re-sampling operations. Departing from previous methodologies, HRSTNet prioritizes high-resolution features and implements multi-resolution feature fusion at each stage. This strategy significantly enhances the interaction among features of varying scales, leading to outstanding segmentation performance within the realm of pure transformer methods.

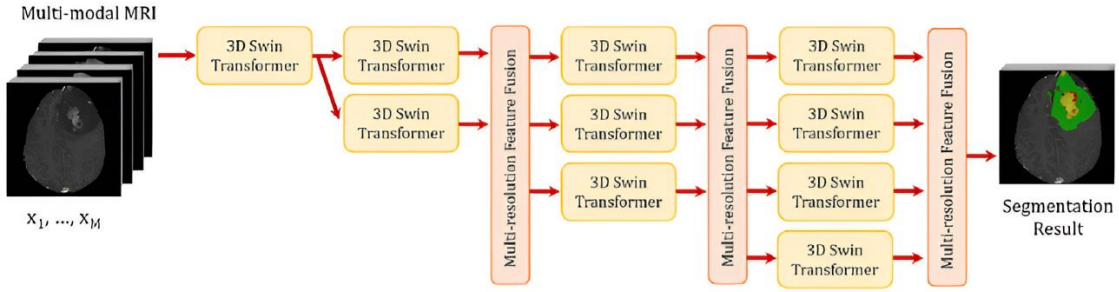

Fig 15. HRSTNet model

In summary, the primary objective of the majority of pure transformer methods is to improve the ability to model local relations while simultaneously reducing model complexity through the introduction of innovative fundamental transformer blocks. Experimental findings consistently indicate that pure transformer methods exhibit competitive performance when compared to state-of-the-art CNNs.

4.2. Hybrid transformer methods

Hybrid transformer methods, in contrast to pure transformer methods, often combine CNNs, known for their ability to capture local feature relations, with transformers, known for their ability to model global feature relations, to improve the accuracy of brain tumor segmentation. These methods are usually divided into three groups based on how CNNs and transformers are combined: methods with a transformer-centric encoder, methods incorporating a transformer between the encoder and decoder, and methods with a transformer-centric decoder (Fig. 16).

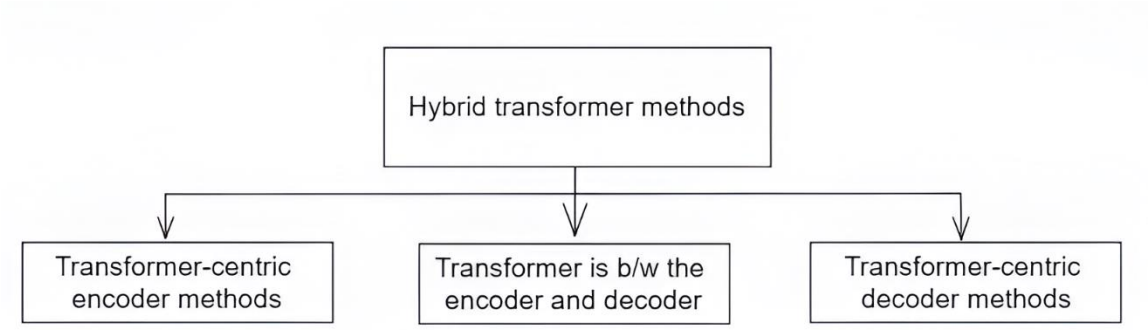
Hybrid transformer methods
Transformer-centric encoder methods
Transformer is b/w the encoder and decoder
Transformer-centric decoder methods

Fig 16. Three groups of Hybrid Transformers

1. Transformer-centric encoder methods

The primary goal of methods with a transformer-centric encoder is to leverage the transformer's capabilities to model extensive dependencies and extract superior encoding features. Consequently, some studies prefer to directly use transformers as encoders.

Recognizing the advantages of SwinT in performance and computational efficiency compared to ViT, many studies prefer to develop encoders based on SwinT. Among them, some researchers directly use 3D SwinT as the encoder and CNN as the decoder to construct a U-shaped network for brain tumor segmentation. A notable example is UNETR, which follows a traditional encoder-decoder architecture (Fig. 17). In UNETR, the encoder architecture is based on ViT, while the decoder consists of several 3x3 convolutions and

2x2 deconvolutions, similar to U-Net. Using only the transformer as the encoder, as opposed to pure transformer methods, can reduce computational overhead without compromising segmentation accuracy.

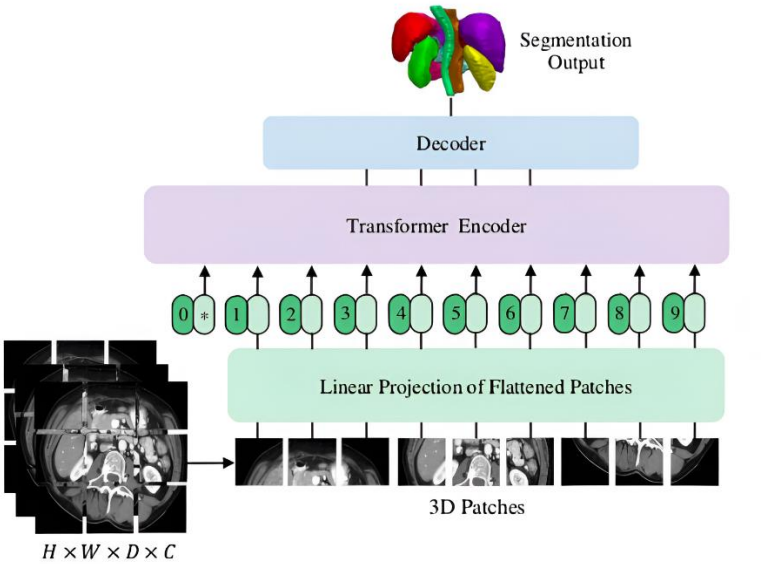

Fig 17. An overview of UNETR reveals that it features a transformer encoder, which directly uses 3D patches, and is connected to a CNN-based decoder through skip connections.

2. Transformer between the encoder and decoder methods

The second category of methods incorporates a transformer between the encoder and decoder. Given that encoding features have lower spatial resolution and richer semantic information, these methods can effectively model global feature relations at reduced computational costs.

Among recent notable developments, the TransBTS model is a pioneering approach that uses a CNN-based encoder to extract features from multi-modal MRI and then segments the encoding features into non-overlapping patches (Fig. 18). After incorporating learnable positional embeddings, several standard transformer blocks are used to capture global relations within the encoding features. Ultimately, by converting the transformer's output into 2D feature maps for input into a CNN-based decoder, final tumor predictions are generated. Despite the limited number of transformer layers, TransBTS has similar parameters to purely CNN-based methods but delivers more satisfactory segmentation results, highlighting the effectiveness of the second category methods.

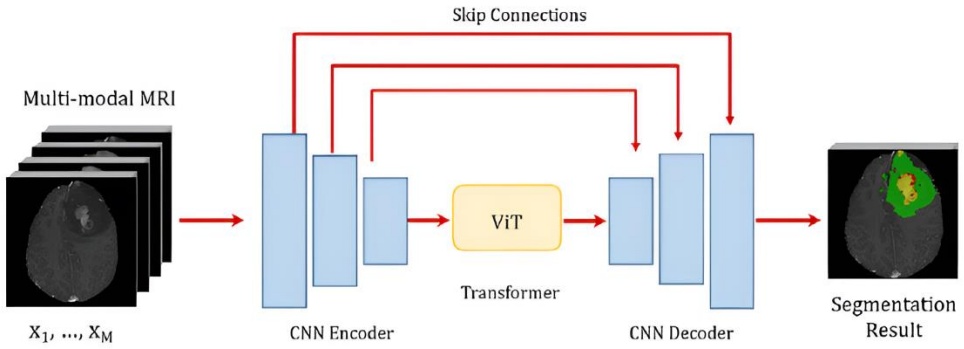

Fig 18. TransBTS model

Another noteworthy model, SegTransVAE, is a multi-task learning framework that includes a shared CNN-based encoder, a transformer-based feature enhancement module, and two specialized CNN-based decoders capable of jointly performing brain tumor segmentation and brain tumor image reconstruction (Fig. 19). Overall, although these methods share similar implementation principles, they illustrate that various established structural innovations, such as attentions and transformer variants, readily enhance performance.

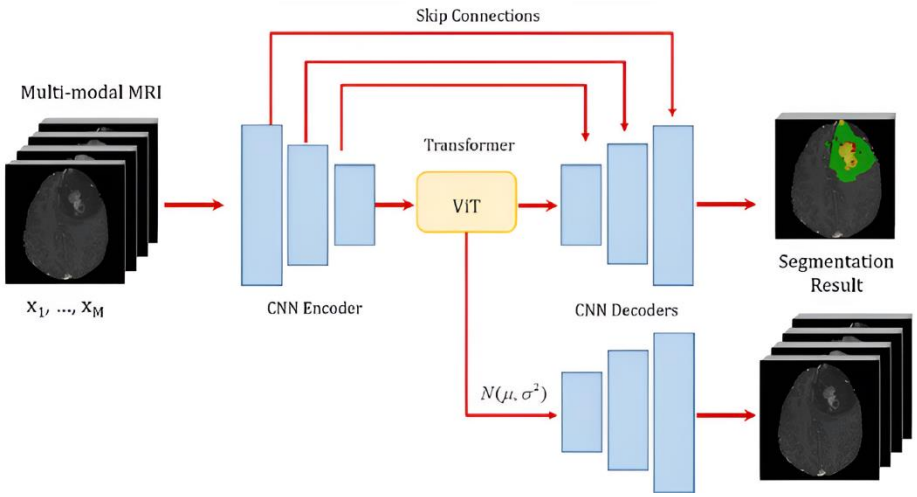

Fig 19. SegTransVAE model

3. Transformer-centric decoder methods

Apart from the aforementioned methods, some models use transformers as decoders for brain tumor segmentation. Recognizing that decoder features embody higher-level semantic information, these approaches utilize the transformer's powerful global modeling capabilities to explore and understand the relationships between tumor regions and the background, thereby enhancing segmentation accuracy.

Here, we finally approach the main model of this thesis, the Multimodal Transformer. The Multimodal Transformer first employs CNN-based encoders to extract features from multi-modal MRI separately for each modality (Fig.20). Then, modality-specific features are transformed into one-dimensional arrays and fed into corresponding transformers. To mitigate the impact of missing modalities, a transformer-based encoder has been proposed

to learn non-local feature relationships between multiple modalities and achieve multi-modal feature fusion. Finally, the fused features are transformed and fed into a CNN-based decoder for the segmentation of incomplete multi-modal brain tumor images. The Multi-modal Transformer comprehensively considers potential inter-modal relationships in various scenarios with missing modalities, thus demonstrating robust performance compared to existing brain tumor segmentation methods.

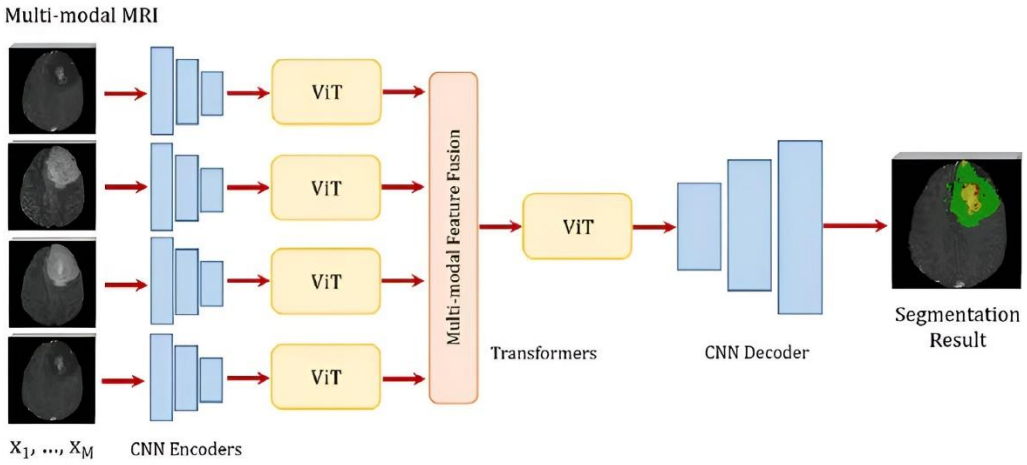

Fig 20. Multimodal transformer

Current research mainly focuses on effectively combining CNNs and transformers to develop hybrid transformer methods. Compared to pure transformer methods, hybrid transformer methods usually better model local feature relationships and have lower model complexity. Experimental data supports the ability of hybrid transformer methods to achieve state-of-the-art segmentation performance. Next, we will discuss the structure of the Multimodal Transformer in detail.

Third Chapter: Multimodal Transformer

Section 1: Method

Let's employ an encoder-decoder architecture to build the multimodal transformer, comprising a hybrid modality-specific encoder for each modality, a modality-correlated encoder, and a convolutional decoder. Additionally, auxiliary regularizers are incorporated into both the encoder and decoder. An overview of multimodal transformer is shown in Fig. 21. The details of each component are elaborated upon in the following sections.

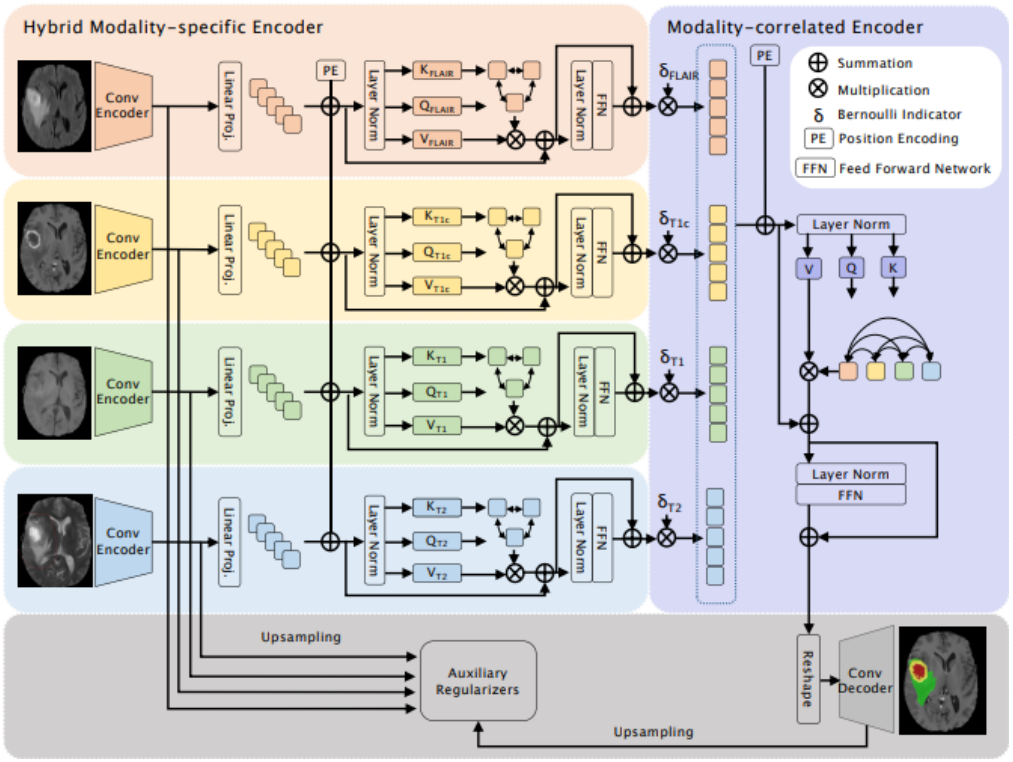

Fig 21. Overview of the multimodal transformer

1.1 Hybrid Modality-Specific Encoder

The goal of the hybrid modality-specific encoder is to extract both local and global content from a particular modality by integrating a convolutional encoder with an intra-modal Transformer. We identify all modalities by M = {FLAIR, T1c, T1, T2}. For an input $X_m \in \mathbb{R}^{1\times D\times H\times W}$ with dimensions D × H × W, m ∈ M, the convolutional encoder is initially used to generate compact feature maps that capture the local context, then it utilizes the intra-modal Transformer to address long-range dependencies globally [17].

1. Convolutional encoder

The convolutional encoder, comprised of several convolutional blocks similar to those in the U-Net encoder, builds feature maps with local context for each modality. The output from the convolutional encoder F_m^{conv} is expressed as:

$$F_m^{local} = F_m^{conv}(X_m;\ \theta_m^{conv})$$

Where $F_m^{local} \in \mathbb{R}^{C\times\left(\frac{D}{2^{l-1}}\right)\times\left(\frac{H}{2^{l-1}}\right)\times\left(\frac{W}{2^{l-1}}\right)}$, C is the channel dimension, and l represents the number of stages in the encoder. Specifically, the encoder is structured as a five-stage cascade, with each stage containing two convolutional blocks. Each block includes cascaded group normalization, ReLU activation, and convolutional layers with a kernel size of 3. At the beginning of the convolutional sequence, the first stage only contains a convolutional layer. Between successive blocks, a convolutional layer with a stride of 2 is used to down sample the feature maps. The encoder stages are equipped with 16, 32, 64, 128, and 256 filters respectively.

2. Intra-modal Transformer

The convolutional encoder, though robust, struggles to establish long-range dependencies essential for accurate brain tumor segmentation. The Intra-modal Transformer is introduced to overcome this limitation, utilizing Multi-head Self Attention (MSA) and a Feed-Forward Network (FFN) to enhance contextual modeling over large distances.

Transformers process embeddings sequentially, converting the feature maps F_m^{local} produced by the convolutional encoder into a 1D sequence, and then into a token space through a linear projection. The flattening of these sequences into tokens compresses spatial information, which is vital for accurate segmentation. To counteract this, a learnable position embedding P_m is introduced, enhancing the features via element-wise summation:

$$F_m^{token} = F_m^{local} W_m + P_m$$

Here, $F_m^{token} \in \mathbb{R}^{C'\times\left(\frac{DHW}{2^{3(l-1)}}\right)}$ represents the tokens, and W_m denotes the weights of linear projection. The MSA establishes intra-modality relationships by examining all possible locations within the feature maps:

$$head_m^i = Attention\left(Q_m^i, K_m^i, V_m^i\right) = softmax\left(\frac{Q_m^i K_m^{iT}}{\sqrt{d_k}}\right) V_m^i$$

$$MSA_m = [head_m^1, \ldots, head_m^N] W_m^o$$

Where, $Q_m^i = LN\left(F_m^{token}\right) W_m^{Qi}$, $K_m^i = LN\left(F_m^{token}\right) W_m^{Ki}$, $V_m^i = LN(F_m^{token}) W_m^{Vi}$, LN denotes layer normalization, d_k represents the dimension of K_m, and N is the number of attention heads (N = 8 here). A perceptron with GELU activation follows, integrating features across the network. The final global context features within each modality are derived from the intra-modal Transformer as follows:

$$F_m^{global} = FFN_m\left(LN(z)\right) + z, z = MSA_m\left(LN(F_m^{token})\right) + F_m^{token}$$

Where $F_m^{global} \in \mathbb{R}^{C'\times\left(\frac{DHW}{2^{3(l-1)}}\right)}$, encapsulating the transformed feature set ready for further processing in the neural network.

1.2 Modality-correlated Encoder

The modality-correlated encoder is crafted to establish long-range correlations across different modalities to obtain modality-invariant features with global semantics pertinent to the tumor region. This is achieved through an inter-modal Transformer.

1. Inter-modal Transformer

Unlike the intra-modal Transformer, the inter-modal Transformer merges embeddings from all modality-specific encoders by concatenation to form the input multimodal token, defined as

$$F^{token} = \left[\delta_{FLAIR} F_{FLAIR}^{global}, \delta_{T1c} F_{T1c}^{global}, \delta_{T1} F_{T1}^{global}, \delta_{T2} F_{T2}^{global} \right] W + P,$$

where $\delta_m \in \{0, 1\}$ represents a Bernoulli indicator that ensures robustness in building long-range dependencies between different modalities, even when some modalities are absent. This type of modality-level dropout is randomly applied during training by setting δ_m to 0. For missing modalities, the multimodal token will be replaced with a zero vector. Subsequently, it is processed by MSD and FFN to produce modality-invariant features across modalities, which is expressed as

$$F^{global} = FFN\left(LN(z)\right) + z, z = MSA\left(LN(F^{token})\right) + F^{token},$$
$$F^{global} \in \mathbb{R}^{C' \times \left(\frac{DHW}{2^{l-1}}\right)}.$$

1.3 Convolutional Decoder

The convolutional decoder aims to progressively restore the spatial resolution from the high-level latent space back to the original mask space. The output sequence F^{global} from the modality-correlated Transformer is reshaped into feature maps matching their original size before flattening. This decoder features a symmetric architecture similar to the U-Net convolutional encoder. Additionally, skip connections between the encoder and decoder are included to retain more low-level details for improved segmentation. Features from convolutional encoders of different modalities at a specific level are concatenated and forwarded as skip features to the convolutional decoder.

1.4 Auxiliary Regularizer

Traditional multimodal learning models often recognize brain tumors by relying on discriminative modalities. These models may suffer significant degradation if the discriminative modalities are absent. Therefore, it is essential to ensure that each convolutional encoder can segment brain tumors independently of other modalities. To achieve this, the outputs of the convolutional encoders are upsampled by a shared-weight decoder to segment tumors from each modality separately. The shared-weight decoder mirrors the architecture of the convolutional decoder. Additionally, auxiliary regularizers are introduced in the convolutional decoder to ensure accurate segmentation even when certain modalities are missing. This is accomplished by interpolating feature maps at each stage of the convolutional decoder to segment tumors through deep supervision. Dice loss

is employed as the regularizer. The overall loss function combines the training loss of the network's output with the auxiliary regularizers, defined as:

$$\mathcal{L} = 1 - Dice = 1 - \frac{2\sum_{c=1}^{C}\sum_{i=1}^{N_c} g_i^c p_i^c}{\sum_{c=1}^{C}\sum_{i=1}^{N_c} g_i^{c2} + \sum_{c=1}^{C}\sum_{i=1}^{N_c} p_i^{c2}}$$

$$\mathcal{L}_{total} = \sum_{i \in M} \mathcal{L}_i^{encoder} + \sum_{i=1}^{l-1} \mathcal{L}_i^{decoder} + \mathcal{L}_{output}$$

Where C is the number of segmentation classes, N_c is the number of voxels of class c, g_i^c is a binary indicator if class label c is the correct classification for pixel i, p_i^c is the corresponding predicted probability, M includes modalities, $M = \{\text{FLAIR}, \text{T1c}, \text{T1}, \text{T2}\}$ and l is the number of stages in the convolutional decoder.

Fourth Part: Model Implementation

Section 1. Dataset and Implementation

In this section, we present the experimental analysis conducted on the Brain Tumor Segmentation Challenge (BraTS) datasets to assess the effectiveness of our approach and to compare it with existing state-of-the-art techniques.

The Brain Tumor Segmentation (BraTS) dataset is a paramount resource for assessing brain tumor segmentation methodologies. It includes a wide range of MRI modalities such as T1, T2, T1c, and FLAIR, as well as manually annotated tumor segmentation masks. The BraTS dataset serves as an invaluable tool for researchers and clinicians engaged in glioma segmentation and brain tumor diagnosis, fostering progress in the field.

The experiments are conducted on the BraTS 2020 dataset, which consists of 369 multi-contrast MRI scans with four modalities: T1, T1c, T2, and FLAIR for training and 125 cases without ground truths for online validation. In the training set, experts have labeled the ground truths of three mutually inclusive tumor regions: whole tumor (WT), tumor core (TC), and enhancing tumor (ET). WT contains all tumor tissues, while TC contains ET, necrotic, and non-enhancing tumor core. All volumes have been co-registered to the same anatomical template and interpolated to the same resolution. To avoid using training subjects for testing, we split the dataset into three parts, and the model is trained for each part without using a pre-trained model. For implementing incomplete modalities, we randomly remove MRI modalities according to a uniform distribution.

The Dice Similarity Coefficient (DSC), defined earlier in the equation, is used for evaluation as it measures the overlap between the predicted results and the ground truths. DSC is particularly useful for MRI brain tumor segmentation because it accurately measures how well the predicted tumor regions match the true annotations. The DSC score is easy to interpret, with values ranging from 0 to 1, where 0 indicates no overlap and 1 indicates perfect overlap, making the metric intuitive. It equally considers false positives and false negatives, providing a balanced assessment of segmentation quality.

The method is implemented in Pytorch using a Nvidia GeForce RTX 3060 GPU. We first resize the MRI data from $155 \times 240 \times 240$ to $128 \times 128 \times 128$ for the non-zero region. Then, we apply z-score normalization for each modality and augment the training set with random cropping, random rotations, and intensity shifts. During the training phase, we set the batch size to 1 and use the Adam optimizer for 200 epochs. The learning rate is initialized at 1e-5 and decays by a factor of 0.5 every 10 epochs. Training the entire network takes about 18 hours.

Section 2. Experiments and Results

1.1 Comparison to Other Methods on Incomplete Multimodal Segmentation

In this section, we compare our method with other state-of-the-art methods based on Dice Scores for incomplete MRI modalities of the BraTS 2020 dataset. For comparison, I selected two previously reviewed models: TransBTS and SegTransVAE. There is a total of 15 possible combinations of available modalities.

In Table 2, we compare the models on the segmentation of the whole tumor and see that our method performs better than TransBTS and SegTransVAE on most subsets with missing modalities. The average results across all subsets are the best. In Table 3, we compare the models on the segmentation of the tumor core, and again the multimodal transformer shows the best average result. In Table 4, we compare the models on the segmentation of the enhancing tumor, and here our model also shows the best result compared to the other two models.

Modalities				Whole tumor		
F	TC1	T1	T2	TransBTS	SegTransVAE	Our
•	○	○	○	71.87	74.31	85.45
○	•	○	○	75.58	76.89	77.20
○	○	•	○	69.18	71.52	**78.16**
○	○	○	•	82.29	86.87	**88.47**
•	•	○	○	83.05	85.21	87.59
•	○	•	○	73.19	80.22	81.33
•	○	○	•	65.37	87.20	88.41
○	•	•	○	82.31	85.51	89.28
○	•	○	•	71.89	87.25	90.05
○	○	•	•	83.88	88.60	89.87
•	•	•	○	84.58	90.60	90.38
•	•	○	•	85.30	89.11	90.32
•	○	•	•	86.53	88.85	90.02
○	•	•	•	86.71	86.03	89.76
•	•	•	•	86.82	89.89	90.61
Average				79.21	84.53	**87.13**

Table 2. Comparison of segmentation results on Whole tumor

In Table 3, we compare the models on the segmentation of the tumor core, and again the multimodal transformer shows the best average result.

Modalities				Tumor core		
F	TC1	T1	T2	TransBTS	SegTransVAE	Our
•	○	○	○	62.93	68.51	**70.65**
○	•	○	○	77.48	81.77	**82.86**
○	○	•	○	60.31	65.91	67.11
○	○	○	•	64.02	70.12	69.31
•	•	○	○	77.60	82.44	84.63
•	○	•	○	78.86	83.01	85.87
•	○	○	•	68.15	71.59	72.27
○	•	•	○	66.20	71.15	73.41
○	•	○	•	68.29	74.14	75.24
○	○	•	•	75.03	81.80	85.53
•	•	•	○	74.15	82.87	86.30
•	•	○	•	78.69	85.07	86.47
•	○	•	•	77.05	84.96	86.92
○	•	•	•	78.58	85.05	87.38
•	•	•	•	79.45	85.49	87.48
Average				72.45	78.25	80.09

Table 3. Comparison of segmentation results on Tumor core

In Table 4, we compare the models on the segmentation of the enhancing tumor, and here our model also shows the best result compared to the other two models.

Modalities				Enhancing Tumor		
F	TC1	T1	T2	TransBTS	SegTransVAE	Our
•	○	○	○	36.56	44.29	43.56
○	•	○	○	67.91	74.85	76.93
○	○	•	○	29.30	38.39	42.60
○	○	○	•	34.68	39.91	43.21
•	•	○	○	71.92	76.01	82.72
•	○	•	○	72.31	78.55	79.20
•	○	○	•	34.78	45.65	48.16
○	•	•	○	44.77	50.87	51.11
○	•	○	•	41.15	49.43	48.42
○	○	•	•	69.12	76.81	81.98
•	•	•	○	42.19	49.82	52.44
•	•	○	•	70.80	77.17	79.45
•	○	•	•	71.56	76.89	78.30
○	•	•	•	72.73	77.76	80.47
•	•	•	•	71.49	78.05	81.58
Average				55.41	62.29	64.60

Table 4. Comparison of segmentation results on Enhancing tumor

Fig. 22 shows that even with only one modality available, the model can achieve accurate brain tumor segmentation. It is observed that the model shows greater improvement when the segmentation target is more complex. These results demonstrate the effectiveness of the multimodal transformer for incomplete multimodal learning in brain tumor segmentation.

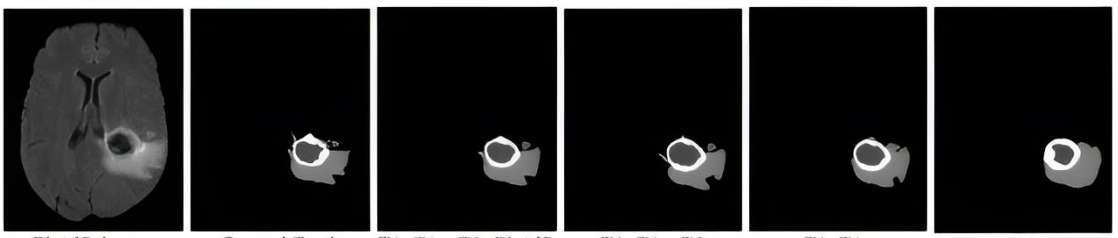

Fig 22. Segmentation results with different available modalities

Fifth Part: Conclusion and Outlook

Section 1. Conclusion

Thus, we have reviewed the main existing methods for brain tumor segmentation, explored how the multimodal transformer works, and identified its advantages. To address the issue of missing modalities in clinical scenarios, the multimodal transformer has proven to be highly reliable. It is particularly noteworthy that the multimodal transformer shows significant performance improvements when more modalities are missing.

The multimodal transformer not only excels in learning features of missing modalities but also effectively integrates multimodal features from complete data, leading to more accurate brain tumor segmentation compared to alternative methods. However, this approach has some limitations. Despite the partial compensation provided by our multimodal transformer for missing features, segmentation performance may still depend on the information from available modalities. Additionally, our network is trained on data with complete modalities to learn features of missing modalities. In real-world hospital datasets, where data may be incomplete, our method can adapt by training on a subset with complete modalities for testing, leveraging the multimodal transformer's ability to model correlations for effective feature fusion even with incomplete data.

Section 2. Outlook

There are several directions for future research and improvements that can be implemented to further enhance the performance and applicability of this approach. One potential improvement, in my opinion, is the use of various regularization methods to boost model performance. Regularization techniques can help mitigate overfitting and improve the model's generalization capabilities, making it more robust to variations in the data. By experimenting with different regularization strategies, we can identify the most effective methods to incorporate into the multimodal transformer architecture.

Additionally, a critical step in the further development of this project is the validation and refinement of the model using real clinical data. While the current results on benchmark datasets, such as BraTS 2020, are promising, applying the model to real-world clinical datasets will provide a more comprehensive evaluation of its practical utility. This transition will help ensure that the model can handle the complexities and variabilities present in actual clinical environments, ultimately leading to more accurate and reliable brain tumor segmentation in medical practice.

By addressing these future directions, we aim to create a more robust and clinically applicable brain tumor segmentation model. The integration of advanced regularization methods and the application to real clinical data will not only improve performance but also enhance the overall impact and usability of the proposed multimodal transformer in healthcare.

References

[1] Wang, Y.P. Medical Image Processing; Tsinghua University Press: Beijing, China, 2012

[2] Kaleem, M.; Sanaullah, M.; Hussain, M.A.; Jaffar, M.A.; Choi, T.S. Segmentation of Brain Tumor Tissue Using Marker Controlled Watershed Transform Method. Commun. Comput. Inf. Sci, 2012

[3] Guo, Y.; Sengur, A. NCM: Neutrosophic C-Means Clustering Algorithm. Pattern Recognit, 2015

[4] Jayanthi, S.; Ranganathan, H.; Palanivelan, M. Segmenting Brain Tumor Regions with Fuzzy Integrated Active Contours. IETE J. Res, 2019

[5] Havaei, M.; Jodoin, P.M.; Larochelle, A.H. Efficient Interactive Brain Tumor Segmentation as Within-Brain KNN Classification. In Proceedings of the International Conference on Pattern Recognition, Stockholm, Sweden, 2014

[6] Vaishnavee, K.; Amshakala, K. An Automated MRI Brain Image Segmentation and Tumor Detection Using SOM-clustering and Proximal Support Vector Machine Classifier. In Proceedings of the 2015 IEEE International Conference on Engineering and Technology, ICETECH, Coimbatore, India, 2015

[7] Selvathi, D.; Selvaraj, H. Segmentation of Brain Tumor Tissues in MR Images Using Multiresolution Transforms and Random Forest Classifier with Ada Boost Technique. In Proceedings of the 2018 26th International Conference on Systems Engineering, ICSEng, Sydney, Australia, 2018

[8] Meier, R.; Bauer, S. Patient-specific Semi-supervised Learning for Postoperative Brain Tumor Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 2019

[9] Saouli, R.; Akil, M.; Kachouri, R. Fully Automatic Brain Tumor Segmentation Using End-to-end Incremental Deep Neural Networks in MRI Images. Comput. Methods Programs Biomed, 2018

[10] Zhao, X.; Wu, Y.; Song, G.; Li, Z.; Zhang, Y.; Fan, Y. A Deep Learning Model Integrating FCNNs and CRFs for Brain Tumor Segmentation. Med. Image Anal, 2017

[11] Badrinarayan, V.; Kendall, A. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell, 2017

[12] Jia, H.Z.; Xia, Y. Learning High-Resolution and Efficient Non-Local Features for Brain Glioma Segmentation in MR Images; Medical Image Computing and Computer Assisted Intervention, MICCAI: Lima, Peru, 2020

[13] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: Proc. International Conference on Learning Representations. ICLR, 2020

[14] Peiris H, Hayat M, Chen Z, Egan G, Harandi M. Hybrid window attention based transformer architecture for brain tumor segmentation. arXiv preprint arXiv, 2022

[15] Jiang Y, Zhang Y, Lin X, et al. A method for 3D multimodal brain tumor segmentation using Swin transformer. Brain Sci, 2022

[16] Li J, Wang W, Chen C, et al. TransBTSV2: towards better and more efficient volumetric segmentation of medical images. arXiv preprint arXiv, 2022

[17] Zhang Y, He N, Yang J, et al. mmformer: multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: Proc. Medical Image Computing and Computer-Assisted Intervention. MICCAI, 2022

Acknowledgements

The four years of study at the University of Science and Technology of China have flown by. During this time, I have learned a great deal, and I would like to thank all the teachers who guided me, the classmates who helped me, and the friends who cared for me.

Firstly, I want to express my gratitude to the University of Science and Technology of China for providing an excellent platform for learning and rich academic resources. The support and education I received at the university have enabled me to develop comprehensively and achieve academic success.

Secondly, I would like to extend my sincere thanks to my supervisor, Professor Xingfu Wang, for his guidance and support during the writing of my thesis. Your professional knowledge and patient guidance have played a crucial role in my research.

Furthermore, I want to thank my class advisor Hailong Wang, as well as all the teachers who taught me throughout my undergraduate studies. Your teachings and care have been invaluable, providing me with an excellent academic environment and a solid foundation of knowledge.

Lastly, I wish to thank my family, classmates, and friends. Thank you for your support and encouragement throughout my studies. Your companionship and assistance have made this time unforgettable.